NOT TO BE FURTHER COPIED OR REPRODUCED

(* BATCHES *)

4

Splitting Up Numbers

(Residuals)

		Exh.		Page
Introduction 4-1		1	a	4-4
4A. Splitting up data 4-2			b	
			c	
4B. Residuals and central values 4-6			d	
residuals, at 4-7		2		4-14/15
central value, at 4-7			e	
- Mean and median 4-8 to 4-10		3		4-19/20
arithmetic mean, at 4-8		4		4-21
- trimean and midmean 4-10a			f	
midmean, at 4-10a		5		4-22c/22d
		6		4/22e/22f
4C. The role of residuals 4-11		7	g	4-26
4D. Clarity of residuals 4-13			h	
resistant, at 4-16		8	k	4-30
4E. Combining residuals across batches 4-18		9		4-31
		10		4-32
4F. A batch and a standard value 4-22a		11		4-33
		12		4-35/36
4G. Choice among middle means 4-22g		13		4-37
**********		14	l	4-40
		15		4-41
4H. How far have we come? 4-27			m	
**********		16	p	4-44
		17		4-46
4K. The effects of changing one value in a batch (background) 4-29		18		4-49
		19		4-50
		20		4-51
4L. Effect on pattern of residuals of choice of central value (background) 4-38		21		4-52
		22		4-53
		23		4-54
4M. Background references 4-42		24		4-55
		25		4-56
**********		26		4-57
4P. Problems for chapter 4 4-43				

4

Splitting Up Numbers
(Residuals)

A detective's work does not consist of a single step. As many stories have told us, he must often clear away layer after layer of irrelevancy or confusion until he can begin work on the crucial layer. As in the detection of crime, so in quantitative detective work.

Stripping away concealment to reveal what is essential may be easy or difficult, may involve stripping away dogs to reveal tigers, or elephants to reveal mice. Sometimes it can be done casually and crudely, at other times, it may require more care and delicacy than we know how to provide. There is no single pattern.

But whatever the pattern, looking below the surface -- without forgetting the surface, which is also often vital -- is an essential of data analysis, perhaps the essential.

Even re-expressing data is an example. By eliminating peculiarities -- perhaps an apparent tendency to stray upward -- that are not essential, re-expression can clear the air, allowing us to see smaller appearances that may be essential.

In this chapter we begin the use of the most powerful tool we know for looking below the surface -- subtraction. By itself, and as a preparer of the way for pictures, subtraction is our most potent tool. It can be used in many ways. This

whole account will not introduce you to all of them. This chapter will, of course, only begin the introduction.

The ideas and attitudes shown here in the simplest case are central to the exploration of data.

4A. Splitting up data

The aim of exploratory data analysis is to "see into" the data. Sometimes one seeks regularities, sometimes irregularities, but one always tries to penetrate the haze of irrelevance and the fog of variability.

One of the greatest single tools in doing this is the separation of the data into parts relevant to separate questions.

Sometimes this separation will require merely throwing pieces of data in the appropriate basket, as when we separate men's weights from women's weights -- or do the same with strengths of grip, lengths of hair, or college board scores. More often, however, this separation will involve tearing apart individual numbers into two or more parts.

Consider an able bowler, who in the last 10 games scored 193, 214, 185, 198, 203, 207, 191, 184, 210, 192. The two most obvious questions about this bowler's performance are these:

- How well does he generally bowl? What score is reasonably central?
- How consistent is his performance?

Let us, then, take apart these 10 scores in some way (or even in alternative ways) which will tend to separate information about these two questions. We shall not seek "the best way" of taking this data apart. (There may well be no "best way", in any realistic sense.) Instead, we will try to select methods which combine understandability with what experience suggests to be reasonable effectiveness.

What can we do to our bowling scores that will not affect what they tell us about our bowler's irregularity or consistency? Surely subtracting the same constant from all the scores will leave their message about variability undistorted.

Subtraction works well more often than not, when we are trying to make this sort of separation. Other devices sometimes do better. Representing the observation as a percentage of some chosen constant -- which is really the same as looking at the ratio of observation to chosen constant -- is often helpful. Because of the last chapter, we know how to turn taking ratios into subtraction; use log values instead of raw values. By doing this we gain both simplicity and use of our full kit of tools. So we do it often.

Exhibit 1 gives a collection of what happens to the 10 bowling scores when we subtract constants. In column (2) we have peeled 100 off each score. Clearly this has left the information about variability undisturbed. Equally clearly, we may as well peel off more. Subtracting 150 from each original score, as in column (3) deserves the same comment. After this subtraction what is left ranges from 34 to 64; clearly we would do well to go further.

The next try, in column (4), is the subtraction of 200. What is left tells us about little except variability. If we want to see variability undisturbed by level, 200 is clearly a more purifying choice than 100 or 150. But what of nearby values? Columns (5) and (6) try out 210 and 190 with

exhibit 1 of chapter 4

The 10 bowling scores, raw and with various consistent shifts

(1) Raw scores

(2) Each score less 100

(3) Each score less 150

(4) Each score less 200

(5) Each score less 210

(6) Each score less 190

(7) Each score less 195

(8) Each score less 199

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
193	93	43	-7	-17	3	-2	-6
214	114	64	14	4	24	19	15
185	85	35	-15	-25	-5	-10	-14
198	98	48	-2	-12	8	3	-1
203	103	53	3	-7	13	8	4
207	107	57	7	-3	17	12	8
191	91	41	-9	-19	1	-4	-8
184	84	34	-16	-26	-6	-11	-15
210	110	60	10	0	20	15	11
192	92	42	-8	-18	2	-3	-7

results that appear somewhat less satisfactory than the subtraction of 200, since the former gives results predominantly made up of negative values, and the latter gives predominantly positive results.

In rather rough terms, we might now answer the two questions as follows:

- The bowler generally rolls about 200 -- give or take an as yet unspecified amount -- this is a central score for him.
- His variability is well illustrated by column (4) of display 1; he is not infrequently 10 or 15 pins above or below his typical performance, rarely more.

This particular batch of 10 bowling scores is unlikely to yield much more information than we have just expressed. As was the case here, a rather free-and-easy approach to the data often squeezes out most of the available "juice". That this is true is important, for it means that when we need to break new ground in exploring our data -- as we often will -- we can often do quite well with a free-and-easy approach.

This does not mean, however, that we necessarily lose by picking out standard (even hide-bound) processes and planning to use them regularly. Even in exploratory data analysis, and even though their use tends to make us somewhat over-rigid, there are often good reasons for choosing and using standard procedures. Two of these reasons stand out:

- Standard procedures reduce effort and increase the chance that an analysis will be made. We had to compute and look at six columns in exhibit 1 in order to

decide that 200 was about a reasonable amount to subtract. Both computation and looking involved work that we might better save for other things if we can. A "once through" calculation would be quicker and easier, and is thus more likely to be undertaken.

- It is much easier to tell people what we have done when this is a standard manipulation. Telling people -- partners, supervisors, colleagues, readers, sometimes even one's self -- what was done is always a very important part of data analysis.

4B. Residuals and central values

We have been splitting up the bowling scores into two parts:

- what we subtracted -- and also thought about for its own sake.

- what was left.

We need to think more about each of these parts -- both in this case and in more general ones -- and to introduce names for them -- again both for this special case and for more general situations.

Our splitting process is two-faced. It points toward what was subtracted. Equally, it points toward what was left. Sometimes our main interest is in the one, sometimes in the other.

A detective looking at a collection of tracks may need to learn that a lame man made them -- that they have a common characteristic. Equally, he may need to learn that they differ among themselves so much that two lame men must have passed that way -- indeed, it may be vital to learn which one passed first.

What is left after subtraction are naturally called

residuals

(Perhaps "left-overs" would have been too informal.) Much of careful, skillful data analysis is spent in making appropriate residuals and then looking hard at them.

As we have seen, we expect residuals to be some positive and some negative. (Otherwise we would feel a better choice should have been subtracted.) Accordingly, it is reasonable to call what has been subtracted a

central value

Often a central value tells us much about the batch with which it is associated.

For a batch of numbers, then, the basic equation is

observation = central value + residual

Depending on our purposes, the important term may be either the central value or the residuals.

What should we choose for a central value? We have already met various summary values -- in particular the median and the trimean. We could use one of these, rather arbitrarily. Let us instead be cautious, and act as if we had not met these summary values. How would we argue, where would we come out, if we started afresh?

* mean and median *

How are we to find a central value in a usefully standardized way? We are working with numbers. The simplest and easiest things to do with them (or with selected ones of them) are to add them up or to order them. (Counting is even simpler, but unless we order them first it is hard to see how to use counting alone to take us along the way we want to go.)

Those who like to add will have noticed that the sum of the numbers in column (4) is not large -- actually -23. They might well want to pick out a number to subtract so that the sum of what is left -- the sum of the "residuals" -- is zero. To do this, of course, they have only to add up the raw values (finding 1977 in the example), count how many are involved (10 in the example), and divide the sum by the count (finding 1977/10 = 197.7 in the example). The result is often called the arithmetic mean, so that one standard choice is

$$\text{arithmetic mean} = \frac{\text{sum}}{\text{count}}$$

(We could have added small numbers instead of large ones had we taken (-23)/(10) = -2.3 as an adjustment to 200, again finding 197.7.)

Others are naturally more concerned with greater or less, and thus with order in general and with + or - in particular. They will have noticed that column (7) has as many + entries as - ones -- five of each. This seems to them a clear sign that 195 is about right. To compute from the raw values, they would arrange the scores in order of size, finding this array:

214
210
207
203
198

193
192
191
185
184

They would naturally pick the middle score in such an ordered list. With 10 scores -- or any even number of scores -- there is no single middle number; rather there is a middle pair -- here 193 and 198.

Subtracting any number that is both greater than 193 and less than 198 will leave five residuals + and five - in this example. To be definite, we might as well go halfway between and use

$$195.5 = \frac{1}{2}(193 + 198)$$

The result, as we have already noted, is usually called the median of the set of numbers. For it we have:

$$\text{median} = \begin{cases} \text{middle value, or} \\ \text{halfway between middle pair} \end{cases}$$

Those who use means as central values tend to think poorly of using medians. At first blush, medians seem to use only one or two values.

Actually, the median does quite well in extracting the juice of the matter. It is usually moderately close to

the mean in juice-extracting power; sometimes it does a very much better job than the mean. (A professional data analyst is often able to infer -- or judge -- which of the two will do better in dealing with a particular sort of data.)

In a situation like the specific example, where the values to be analyzed seem to behave themselves fairly well, it makes little difference whether we use the mean or the median as the typical value. But suppose our bowler had been "hot" when he rolled the ninth of these games, scoring 300 rather than 210. This change would raise the mean of the ten games by 9, but it would not change the median at all.

Sometimes we do better to regard "wild shots" or outliers -- values such as a 300 would be for a bowler who usually rolls near 200 -- as the sort of accidents that should be kept from influencing our central value -- here our central score for this bowler. In particular, errors of observation, recording, or copying may be a major cause of such stray values. If the ninth game had come to us as 400 or 310 we would have known that an error was involved. (The maximum possible score in bowling is 300.) When it is 300 we can wonder about an error.

When we feel that widely straying values are likely, the median is almost always a better choice than the mean. Even when we only fear values that stray much less far -- that are unusual but neither errors nor exceptions -- the median is likely to be the better of these two choices.

* trimean and midmean *

Let us now recall our earlier chapters, and ask about other summary values, and how they are likely to serve as central values.

The median was important to us, but, at least for larger batches, we gave even more attention to the trimean. There seems to be no reason not to expect trimeans to be good central values.

In section 2K, which was optional, we looked at midmeans defined as

midmean = mean of values reached between the hinges.

We can expect the midmean to be useful, also, although it often involves more arithmetic effort, than the trimean.

Both trimean and midmean, then, are candidates whose performance we may wish to look at closely.

4C. The role of residuals

There are many uses for residuals. Indeed, we shall give more attention to the various kinds of residuals that we are going to meet than to any other aspect of our data.

In this chapter, residuals are defined by

observation = central value + residual

In later chapters, we shall need a more general definition, one that we can use now to understand residuals and their uses better. The analysis of data almost always involves two closely related threads:

- focusing of attention, in which we sort out certain aspects of the data -- certain directions from which to view it -- that seem more likely to be helpful.

- summarization, in which we try to make explicit an incomplete description of the data, perhaps by one number, three numbers, eleven numbers, or a formula with numerical coefficients.

Good summarization ought to be guided by wise focusing of attention. Standard ways of summarization should reflect aspects of the data that are rather frequently useful.

The descriptions that are offered by a summarization -- be it good or bad -- are always incomplete. They must be, for nothing that contains all the detail can truly be a summary. In particular, then, our incomplete descriptions will tell us something -- but not all -- about each of our observed values, just as a central value for a batch tells us something, but not all, about the batch's individual values.

The general equation is

observation = incomplete description + residual

The residual tells us about the corresponding observation, over and beyond what the incomplete description tells us. We want it to do this clearly and effectively. Whether or not this happens will depend on the behavior of the incomplete description. The residual does not exist until the incomplete description is fixed.

Since the incomplete description is usually found from the observations -- we are now used to central values found from batches as one example -- they are often said to be fitted. Thus compactness and a useful reminder are provided by writing the basic equation as

observation = fit + residual

There are many uses for residuals -- for what is left of our observations after an incomplete description, usually fitted to the data, has been removed.

The kind of residuals we are now discussing, observations minus some central value, are only the simplest instance of a very general concept that is broadly useful.

Almost every type and sort of data analysis makes some use of some sort of residuals.

4D. Clarity of residuals

Exhibit 2 sets out the residuals after fitting each of four central values -- (i) the mean, (ii) the median, (iii) the midmean, (iv) the trimean -- to two batches, once to our original bowling scores and once to the same scores with one score changed -- 203 replaced by 1203.

Comparing the two columns of residuals from the mean (panel B) shows us that every single residual from the mean has been greatly changed. This had to happen because the extreme change of one value has violently changed the mean. If we want residuals to tell us about corresponding observed values, rather than about some mixture of all observed values, this is far from being satisfactory.

Comparing the two columns of residuals from the median shows us that, because this particular extreme change of one value did NOT change the median at all, only one residual from the median was changed -- the one that warns us about 1203 as an observed value, the one that we want to show a change.

exhibit 2 of chapter 4

Residuals from various central values before and after modification of one value

A) ASSUMED OBSERVATIONS and SOME CENTRAL VALUES

in original order		in increasing order	
original	modified	original	modified
193	193	184	184
214	214	185	185
185	185	191	191
198	198	192	192
203	1203	193	193
207	207	198	198
191	191	203	207
184	184	207	210
210	210	210	214
192	192	214	1203
	mean	197.7	297.7
	median	195.5	195.5
	midmean	196.5	197.5
	trimean	197.2	198
	rounded mean	198	298
	rounded median	196	196
	rounded midmean	196	198
	rounded trimean	197	198

(Cont'd)

exhibit 2 (Cont'd)

B) RESIDUALS -- from various (rounded) central values, entries in original order

From means		From medians	
original	modified	original	modified
-5	-105	-3	-3
16	-84	18	18
-13	-113	-11	-11
0	-100	2	2
5	905	7	1007
9	-91	11	11
-7	-107	-5	-5
-14	-114	-12	-12
12	-88	14	14
-6	-106	-4	-4

From midmean		From trimeans	
original	modified	original	modified
-3	-5	-4	-5
18	16	17	16
-11	-13	-12	-13
2	0	1	0
7	1005	6	1005
11	9	10	9
-5	-7	-6	-7
-12	-14	-13	-14
14	12	13	12
-4	-6	-5	-6

Similarly, comparing the pairs of columns of residuals from the midmean and the trimean shows us that, because midmean and trimean were only slightly affected, the residuals from them were hardly changed at all, except for the one that needs to change greatly to warn us about the single strongly changed observation.

This display illustrates a general fact: residuals from the median or midmean or trimean are much more likely to be helpful than residuals from the mean -- they tend to paint a clearer picture -- they are much less likely to be confused by the effects of a few wide-straying values.

To put it another way, the trimean, the midmean, and the median are

resistant

to the effects of a single wide-straying value -- or of a few such values -- much more resistant than the mean.

For useful residuals, then, we shall tend to avoid calculating residuals from the mean.

Because it is a good compromise, generally useful for all purposes, we will plan to use a **trimean or midmean as** our central value of choice in forming residuals.

Except for small batches, we will also use one of these most of the time as a central value for summarization, although we must recognize that there are circumstances where the mean is the appropriate summary -- usually because we are really concerned with totals -- and circumstances where the median is the convenient summary -- frequently either because re-expression of the median is easy or because the batch is small.

There is no reason for forced consistency. Using different central values for different purposes in analyzing even a single batch of values is not an error. Coat and trousers need not always match. Suppose an observation to be 43, while the mean of the batch to which it belongs is 37 and the midmean is 34, so that we have the pair of decompositions:

$$43 = 37 + 6 \quad \text{(mean + residual)}$$
$$43 = 34 + 9 \quad \text{(midmean + residual)}$$

There can be circumstances where we would want to use 37 as the central value (perhaps because totals are important) and 9 as the residual (probably because we would like our residuals resistant to wild values of other observations). Such a procedure is not inconsistent, as long as we are aware, both of what is being done and what may be essential for effective analysis.

Apparent consistency in exploratory data analysis is as likely to be evil as to be good. As in all exploratory fields, flexibility is usually a mark of wisdom.

4E. Combining residuals across batches

Now that we are well equipped to find residuals, one for each observation in a batch, we are in a position to combine the residuals from several related batches. This often allows us to see deeper into our data.

Exhibit 3 starts with the mean monthly temperatures in New York City for each of the 12 months in seven selected years about 10 years apart from 1871 to 1941. We all expect summer months to be warmer than winter months, so the natural approach to this data is to treat it as 12 batches -- one for each month -- each of seven observations. Residuals from midmeans are calculated in panels B and C, and shown in stem-and-leaf form in panel D.

The most noticeable feature of the combined residuals is their tendency to straggle at the ends. The ends of at least five months straggle upward, and of at least three straggle downward. It seems easier for a month to be unusually warm than for it to be unusually cold.

When we turn back to panel C we see a tendency for the large residuals to come in the winter. (The three residuals outside ±8°F are 2 Januaries and 1 March.) Accordingly it is natural to divide the 12 months into 6 cold months -- midmeans up to 50° -- and 6 warm months -- midmeans above 55°. This is done in exhibit 4.

exhibit 3 of chapter 4

Mean monthly temperatures in New York City in selected years

	Jan	Feb	Mar	Apr	May	Jun	July	Aug	Sept	Oct	Nov	Dec
A) MONTHLY MEAN TEMPERATURES - °F												
1871	30.4	31.8	43.6	53.6	60.8	69.1	71.9	73.0	60.8	54.9	39.3	29.7
1881	25.8	29.5	36.9	46.0	60.2	64.2	72.6	73.1	72.2	59.1	46.3	40.7
1890	40.2	40.4	37.5	47.6	58.8	68.7	70.1	71.5	69.6	56.1	43.2	34.6
1903	30.6	34.4	47.5	52.2	64.1	64.0	74.2	69.2	65.4	56.6	41.4	30.1
1923	31.0	27.1	37.0	49.4	59.3	72.0	72.3	71.3	67.0	55.6	45.4	42.0
1932	42.8	36.0	37.0	48.4	60.9	69.0	73.9	74.8	67.7	57.6	43.7	39.0
1941	29.8	31.4	35.7	56.0	63.0	70.2	74.7	72.7	69.0	60.0	49.6	38.2
B) ORDERED within MONTH												
	42.8	40.4	47.5	56.0	64.1	72.0	74.7	74.8	72.2	60.0	49.6	42.0
Hinge	40.2	36.0	43.6	53.6	63.0	70.2	74.2	73.1	69.6	59.1	46.3	40.7
	31.0	34.4	37.5	52.2	60.9	69.1	73.9	73.0	69.0	57.6	45.4	39.0
	30.6	31.8	37.0	49.4	60.8	69.0	72.6	72.7	67.7	56.6	43.7	38.2
	30.4	31.4	37.0	48.4	60.2	68.7	72.3	71.5	67.0	56.1	43.2	34.6
Hinge	29.8	29.5	36.9	47.6	59.3	64.2	71.9	71.3	65.4	55.6	41.4	30.1
	25.8	27.1	35.7	46.0	58.8	64.0	70.1	69.2	60.8	54.9	39.3	29.7
Midmean	**30.7**	**32.5**	**37.2**	**50.0**	**60.6**	**68.9**	**72.9**	**72.4**	**67.9**	**56.8**	**44.1**	**37.2**
C) RESIDUALS -- also ordered												
	12.1	7.9	10.3	6.0	3.5	3.1	1.8	2.4	4.3	3.2	5.5	4.7
	9.5	3.5	5.4	3.6	2.4	1.3	1.3	0.7	1.7	2.3	2.2	3.4
	0.3	1.9	+0.3	2.2	0.3	0.2	1.0	0.6	1.1	0.8	1.3	1.7
	-0.1	-0.7	-0.2	-0.6	0.2	0.1	-0.3	0.3	-0.2	-0.2	-0.4	0.2
	-0.3	-1.1	-0.2	-1.6	-0.4	-0.2	-0.6	-0.9	-0.9	-0.7	-0.9	-2.7
	-0.9	-3.0	-0.3	-2.4	-1.3	-4.7	-1.0	-1.1	-2.5	-1.2	-2.7	-7.2
	-4.9	-5.4	-1.5	-4.0	-1.8	-4.9	-2.8	-3.2	-7.1	-1.9	-4.8	-7.6

(Cont'd)

exhibit 3 (Cont'd)

D) STEM-and-LEAF -- for all 12 batches of residuals combined, unit = 0.1°F.

1	*1	12	1
.		11	
2		10	3
3		9	5
.		8	
4		7	9
5		6	04
7		5	5
9		4	37
15		3	565124
20		2	24432
29	H8.5	1	938307137
to 40		0	333221763819
✓ 19	M2	-0	1397223642369292749
to 25	H1	-1	156380129
16		-2	48577
11		-3	02
9		-4	90798
4		-5	4
.		-6	
3	*5	-7	126

[84,*HMH*: -7.6, -1.2, -0.2, 1.8 12.1]

E) SOURCE

The World Almanac and Book of Facts for 1958, p. 435.

Their source: U. S. Weather Bureau, New York City.

exhibit 4 of chapter 4

The Stem-and-leaf of panel D of exhibit 12 separated into cold months and warm months

m

A) STEM-and-LEAVES for 6 MONTH BLOCKS

Nov-Apr

to 1	12	1
.	11	
2	10	3
3	9	5
.	8	
4	7	9
5	6	04
7	5	5
8	4	7
12	3	456
14	2	22
16	1	793
to 19	0	933
✓ 10	-0	4913972236
to 13	-1	156
10	-2	764
7	-3	0
6	-4	890
3	-5	4
	-6	
2	-7	26

May-Oct

to 12	12	
11	11	
10	10	
9	9	
8	8	
7	7	
6	6	
5	5	
4	4	3
1	3	512
4	2	443
7	1	383071
13	0	32217638
to 21	-0	423692927
✓ 9	-1	380129
to 12	-2	85
6	-3	2
4	-4	79
3	-5	
.	-6	
1	-7	1

[42, *HMH*: -7.6,-1.6,-0.2,3.4,12.1] [42, *HMH*: -7.1,-1.1,0.0,1.3,4.3]

We now see that the straggling is very much broader in the cold months, although we have to admit that it is present in both six-month segments. If we were prepared to start with data for more years, it would be natural to divide, not into six-month blocks, but into blocks of three or two months, or even to work with single months.

4F. A batch and a standard value

Particularly when we are studying new measurement methods, we often have batches of numbers with each of which goes a standard value. What we often want to look at in such a situation is both the

deviation = central value - standard value

and the

residuals = observations - central value .

If we have several batches usually corresponding to alternative materials, people, or situations, we are likely to want to combine both the deviations and the residuals.

Exhibit 5 shows results of Geissman Kaufman and Dollman on traces of acetylene in air, and carries through the analysis.

We have combined the deviations and, separately, the residuals, in exhibit 3 although a hard look shows us that we should not be happy with this analysis. For the three samples with batches of about 12, the trimeans and differences of hinges were:

tri:	10	50	157
dh:	1	4h	7h

Clearly the spreads within a batch would be more constant if we went to logs.

Exhibit 6 shows the result of going over to logs. The difference of hinges for the three large samples are now .04, .04, and .03 which appear pleasantly constant.

The results of this analysis can, then, be reasonably well summarized by either set of figures in panel E. In general, then, these traces of acetylene were assessed quite fairly (the median deviation could hardly be smaller) reasonably accurately (the midspread of deviations is 3-5%, the range 7%) and, except for a few strays, reasonably precisely (the midspread of the residuals is 5%, although the range is 40%).

exhibit 5 of chapter 4

Determination of traces of acetylene in air

(parts per ten million)

A) The DATA

Sample	Standard Value	Observed Values	Residuals*
A	10	0|99 1|011001100	-0|11 0|011001100
B	40	4|1121	0|0010
C	50	4|889959 5|02841 6|1	-0|221151 0|02841 1|1
D	150	14|99 15|8751963 16|21703	-0|882616 0|1025436 1|0

* From trimeans

B) CALCULATION of TRIMEANS

Sample	Count	Hinge	Median	Hinge	Trimean	Std. Value
A	11	10	10	11	10	10
B	4	41	41	41h	41	40
C	12	48h	49h	53	50	50
D	14	153	157h	160h	157	150

C) The DEVIATIONS COMBINED

		H	0
0	0100	M	0h
0	7	H	4
		tri	1
		dh	4

D) The RESIDUALS COMBINED

			(41)
-0	58866	H	-1
-0	112211121	M	0
0	011001100001002411024 3	H	1
0	856	tri	0
1	10	dh	2

(cont'd.)

exhibit 5 (cont'd.)

E) The SOURCE

T. A. Geissman, S. Kaufman, and D. Y. Dollman 1946. The determination of traces of acetylene in air. Analytical Chemistry 19: 919-921. Table I on page 291.

3Ih

exhibit 6 of chapter 4

The acetylene determination in logs

(logs of parts per ten million)

A) The DATA

Sample	Standard Value	Observed Values	Residuals*
A	1.00	9|55 10|044004400	-0|6611111 0|033003300
B	1.60	16|1121	0|0010
C	1.70	16|889959 17|026319	-0|221151 0|026319
D	2.18	21|779898 22|0001120	-0|331212 0|00011201

* From trimeans

B) CALCULATION of TRIMEANS

Sample	Count	Hinge	Median	Hinge	Trimean	Std. Value
A	11	1.00	1.00	1.04	1.01	1.00
B	4	1.61	1.61	1.61h	1.61	1.60
C	12	1.68h	1.69h	1.72h	1.70	1.70
D	14	2.18	2.20	2.21	2.20	2.18

C) The DEVIATIONS COMBINED -- unit = 0.01 in log

		(4)	
-0	1	H	-.00h
0	002	M	.00
		H	.01
		tri	.00
		dh	.01h

D) The RESIDUALS COMBINED -- unit = 0.01 in log

		(41)	
-0	665	H	-.01
-0	1111122111331212	M	.00
0	33330010023100011201	H	.01
0	69	tri	.00
		dh	.02

(cont'd.)

exhibit 6 (cont'd.)

E) SIMPLE SUMMARIES

in logs

median of deviations = 0.00
midspread of deviations = 0.01h
range of deviations = 0.03
midspread of residuals = 0.02
range of residuals = 0.15

in %

median of deviations = 0 %
midspread of deviations = 3.5%
range of deviations = 7%
midspread of residuals = 5%
range of residuals = 40%

Examples (see exhibit 3)

log 1.00 = .00
log 1.035 = .01h
log 1.07 = .03
log 1.05 = .02
log 1.40 = .15

4G. Choice among middle means

In a later chapter, we will meet modified forms of trimean and midmean. At the end of this section we will have a few words about exact midmeans. For the present, though, our choice is between trimeans and midmeans.

The grounds for choice depend upon the size of the batch. In terms yet to be explained, we can put it thus: In large batches we balance "ease of calculation" against "coarseness of result". In small batches we balance "resistance" against "grasp".

To calculate the midmean of a large batch, we must:

- find the hinges.
- isolate the numbers ranked between the hinges.
- add up the isolated numbers.

For the worker with pencil and paper -- thanks to stem-and-leaf to help with the other steps -- the last step is the tedious one. Even with the techniques to be explained in chapter 7z, the effort required easily becomes unfortunately large.

When the numbers of a large batch are repeated many times, as when the batch gives the number of living children in each of 9,739 families, the midmean can take many values, very close to one another. The trimean is much more restricted. When dealing with whole numbers, as with numbers of children, its values must be multiples of 1/4. These are separated by at least 1/4. In such circumstances, the answer given by the trimean is "coarse".

Unless we run into difficulty with coarseness -- which can be avoided another way, usually easier, as chapter 7x will explain -- we expect to use trimeans as summaries or central values of large batches.

In small batches, the considerations are different. We ordinarily want a central value that will be resistant to the effects of at least one wildly straying value.

For batches of 1 or 2, we have no choice. All reasonable central values are the same.

For batches of 3 or 4, there is a choice, and a desire for resistance settles it. The trimean involves all the observations. The median -- which is the midmean for batches of 3 or 4 -- involves only the center one or two observations. For resistance we must choose it.

For batches of 5 or 6, there is a choice, with advantages on both sides. The median -- still the same as

the midmean -- avoids two observations at each end, and is thus even more resistant than the trimean, which avoids only one at each end. Conversely, the trimean takes account of more of the observations. Different choices here are reasonable. We shall not attempt to make a binding recommendation at this point.

For batches of 7 to 9 observations, a little less arithmetic is required for the midmean than for the trimean. Since the median involves still less arithmetic, we usually use either trimean or median.

The overall conclusion is clear, but not quite simple:

- so long as we avoid difficulties with coarseness, we are likely to want to use trimeans for large batches and medians for small ones. The exact change-over point can often be an individual matter, coming anywhere from between 4 and 5 to between 10 and 11.

* the exact midmean *

It can sometimes be convenient, when programming a computer, to set aside n/4 observations at each end -- and form the mean of the remaining n/2 -- where there are n observations in a batch. This procedure requires taking 1/4's 1/2's and 3/4's of observations into our midmean, but otherwise uses simple formulas which can be used, without modification, for batches of even one or two observations.

For three observations, where n/4 = 3/4, we would find, in particular

$$\text{exact midmean} = \frac{\frac{1}{4}\,(\text{lowest}) + (\text{middle}) + \frac{1}{4}\,(\text{highest})}{\frac{1}{4} + 1 + \frac{1}{4}}$$

Because the extreme observations are used, even if only with the small coefficient 1/4, the exact midmean in batches of three cannot possibly be as resistant as the median, which is the midmean. Our initial choice to prefer the midmean to the exact midmean is based in part on this consideration (batches of three are far from being uncommon), and in part in the greater ease, for paper and pencil, of giving all the observations that contribute directly to the midmean the same coefficient, namely "one".

We would feel free to use the exact midmean in place of the midmean for batches of four or more anything this made the calculation easier.

It may be illuminating to compare how the various middle means weight the ordered observations in small batches. We can always use whole number weights, as when -- for batches of 3:

$$\text{exact midmean} = \frac{\frac{1}{4}\,(\text{lowest}) + (\text{middle}) + \frac{1}{4}\,(\text{highest})}{1\frac{1}{2}}$$

$$= \frac{1(\text{lowest}) + 4(\text{middle}) + 1(\text{highest})}{6}$$

and

$$\text{trimean} = \frac{1(\text{lowest}) + 2(\text{middle}) + 1(\text{highest})}{4}$$

Exhibit 7 sets out the comparison in schematic form. We see that the exact midmean should probably be thought of as "between" the trimean and the midmean.

exhibit 7 of chapter 4

Schematic comparison of the weights with which ordered observations appear in various middle means
(Batches of 3 to 7 and 9 to 10*)

A) BATCHES of 3 to 6

trimean	midmean (= median)	exact midmean
x 1	x	x 1
x 2	x 1	x 4
x 1	x	x 1
x 1	x	x
x 3	x 1	x 1
x 3	x 1	x 1
x 1	x	x
x	x	x
x 1	x	x 3
x 2	x 1	x 4
x 1	x	x 3
x	x	x
x	x	x
x 1	x	x 1
x 1	x 1	x 2
x 1	x 1	x 2
x 1	x	x 1
x	x	x

B) BATCHES of 7, 9 and 10

trimean	midmean (≠ median)	exact midmean
x	x	x
x 1	x	x 1
x	x 1	x 4
x 2	x 1	x 4
x	x 1	x 4
x 1	x	x 1
x	x	x
x	x	x
x	x	x
x 1	x	x 3
x	x 1	x 4
x 2	x 1	x 4
x	x 1	x 4
x 1	x	x 3
x	x	x
x	x	x
x	x	x
x	x	x
x 1	x	x 1
x	x 1	x 2
x 1	x 1	x 2
x 1	x 1	x 2
x	x 1	x 2
x 1	x	x 1
x	x	x
x	x	x

C) NOTES

* For batches of 2 all these middle means are equal. For batches of 8 the two midmeans are equal.
The x's indicate the ordered values of batches of the indicated sizes. The 1,2,1 and 1,4,1 patterns for batches of 3 correspond to the examples treated in text.

*****

4H. <u>How far have we come?</u>

In this chapter we have begun to split up numbers, not yet according to the general

observation = fit + residual

but according to its special case

observation = central value + residual

After looking at means and medians, we have agreed to take middle means as our standard choice for central values. There can be good reasons for selecting either means or medians in special cases, but middle means come far closer to being a generally useful compromise.

The trimean -- the mean of the hinges (each counted once) and the median (counted twice) -- are both very useful. Paper and pencil work, even on large batches, is easy for either. (On a computer, we may want to use the midmean or the exact midmean.)

In this chapter, the major consideration in choosing a central value has been what its use does to residuals. Middle means -- midmeans or trimeans -- make the resulting residuals do quite a good job of being mainly responsive to the value of the observation from which they come and reasonably resistant to the values of the other observations.

The median, as a mean of one or two of the middle observations, shares this property of resistance. (In fact, it is even more resistant.) Since it uses only the most central observations, however, it does not behave like a middle mean for batches of 7 or more. (For batches of 6 or fewer, the median is the midmean.)

We have spent considerable time in this chapter on the reasons for, and notion of, a middle mean and on the definition and easy calculation of various kinds of middle means. Doing BOTH these things is important. It is important to understand why one might want a middle mean -- why it might serve us well. It is also important that, when they would serve us well, middle means be used. As Churchill Eisenhart has succinctly put it -- in a more technical, but only slightly different context:

> "The practical power of a statistical test is the product of two numbers: the mathematical power of the test and the probability that the test will be used."

In practice, neither effectiveness nor frequent use is enough alone. We need both. Paper and pencil work is well served by the median for small batches and the trimean for large batches. The cross-over can come as low as between 4 and 5 or as high as between 10 and 11.

So far as arithmetic goes, the trimean shows us that with stem-and-leaf to order our values we can keep our additionand division simple. The need to form residuals, however, makes it clear that there is no substitute for fast and easy subtraction.

We have also looked at the case where a batch comes with a standard value, and there are likely to be several batches. When we want to combine across batches, re-expression may be called for.

*****

4K. <u>The effects of changing one value in a batch</u> (background)

We have stressed the fact that median and midmean are much more resistant to the straying of a single value -- or a few values -- than is the mean. A few detailed examples can help us to understand a little more about both what has to happen and what does happen.

It will be convenient to work with batches of modest size, and it is nice to base the example on a classical set of data. Exhibit 8 shows the effect of changing the first (by date, see exhibit 16 for problem 4) of Lord Rayleigh's seven weights of nitrogen from air to each of a number of other possible values. The mean shifts steadily, and can clearly be forced to move arbitrarily far. As the one changed value moves through a particular short interval, the median moves rapidly. Otherwise, the median holds to one of two values. The midmean behaves similarly, moving less rapidly over a somewhat longer interval. The trimean behaves rather similarly, moving in three spurts

Exhibit 9 shows the behavior of mean and median graphically. Clearly the median responds much less vigorously to very stray values. Exhibit 10 shows the behavior of the midmean against a background of mean and median. So far as moving this particular value in this particular batch goes, the midmean clearly has advantages. Exhibit 11 shows the behavior of the trimean against a background of mean and median. It, too, has clear advantages for this particular batch and moving value.

exhibit 8 of chapter 4

The effects of changing a single value in Lord Rayleigh's seven weights of nitrogen from air

A) VALUES in TIME ORDER

modified				original	modified		
2.30977	**2.30987**	**2.30997**	**2.31007**	**2.31017**	**2.31027**	**2.31037**	**2.31047**
2.30986	2.30986	2.30986	2.30986	2.30986	2.30986	2.30986	2.30986
2.31010	2.31010	2.31010	2.31010	2.31010	2.31010	2.31010	2.31010
2.31001	2.31001	2.31001	2.31001	2.31001	2.31001	2.31001	2.31001
2.31024	2.31024	2.31024	2.31024	2.31024	2.31024	2.31024	2.31024
2.31030	2.31030	2.31030	2.31030	2.31030	2.31030	2.31030	2.31030
2.31028	2.31028	2.31028	2.31028	2.31028	2.31028	2.31028	2.31028

B) VALUES REORDERED -- unit .00001, referred to 2.31000

-23	-14	-14	-14	-14	-14	-14	-14
-14	**-13**	**-3**	1	1	1	1	1
1	1	1	**7**	10	10	10	10
10	10	10	10	**17**	24	24	24
24	24	24	24	24	**27**	28	28
28	28	28	28	28	28	30	30
30	30	30	30	30	30	**37**	**47**

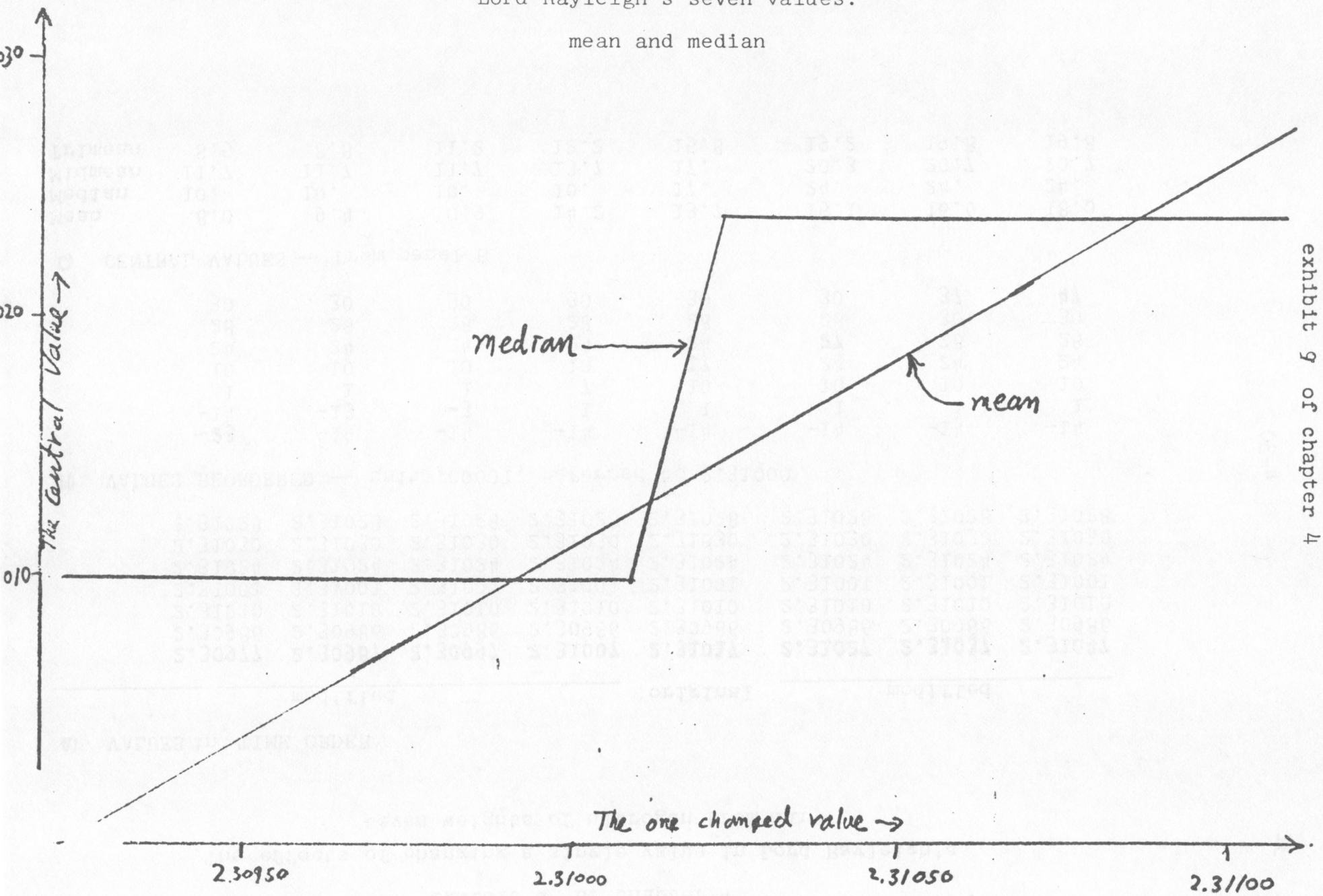
C) CENTRAL VALUES -- from panel B

Mean	8.0	9.4	10.9	14.2	13.7	15.1	16.6	18.0
Median	10.	10.	10.	10.	17.	24.	24.	24.
Midmean	11.7	11.7	11.7	13.7	17.	20.3	20.7	20.7
Trimean	8.5	8.8	11.2	12.2	15.8	19.2	19.8	19.8

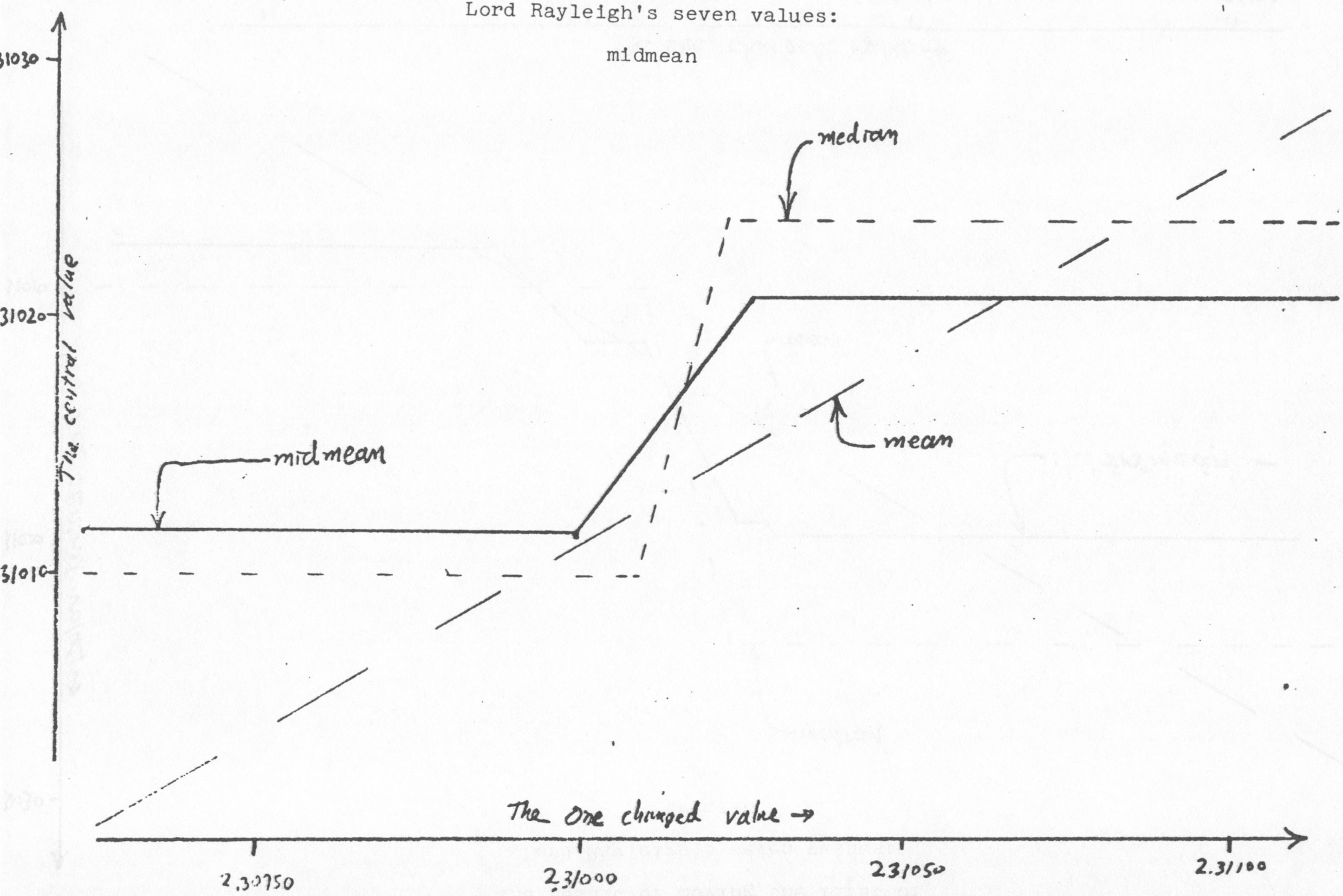
The result of moving the first of

Lord Rayleigh's seven values:

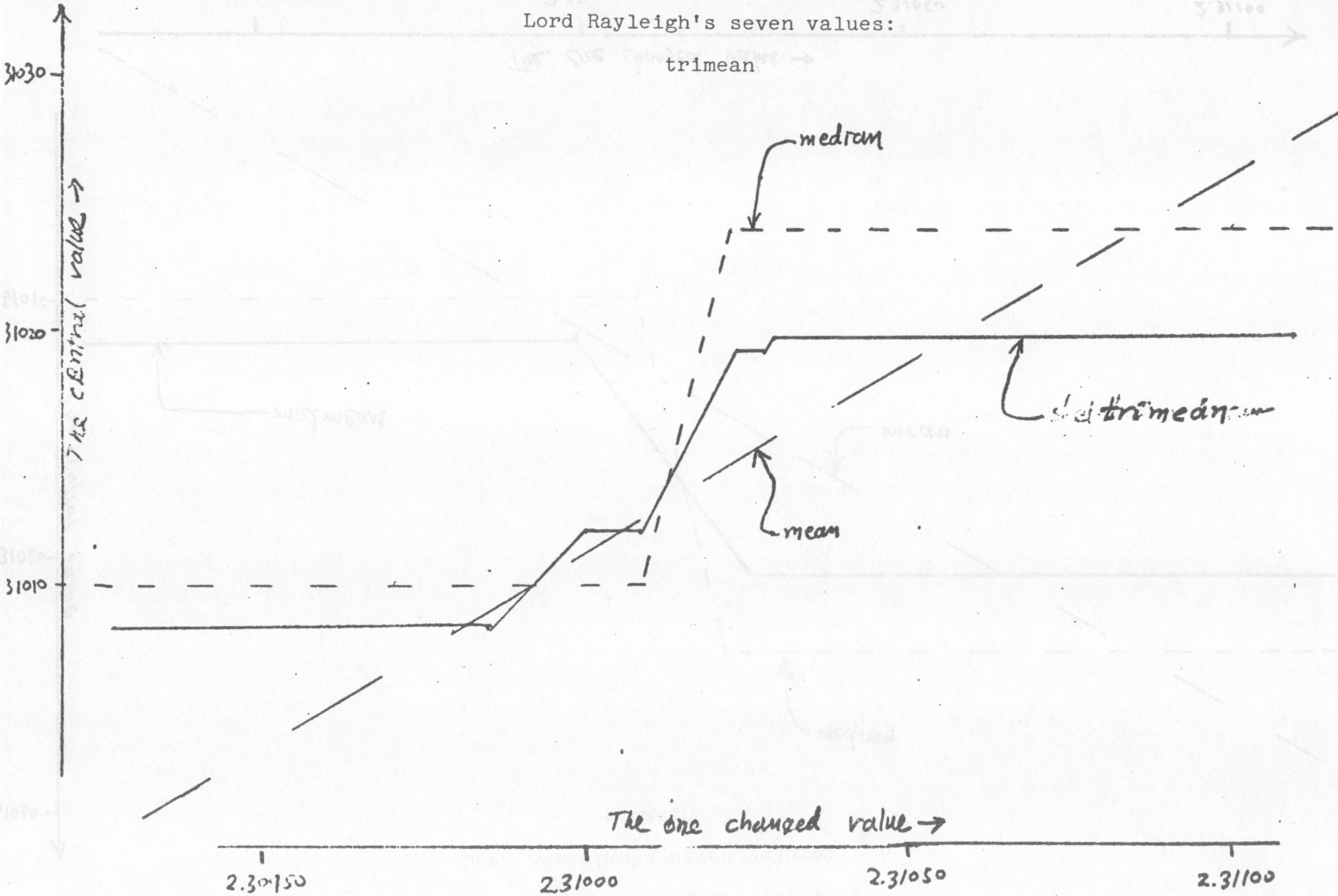
mean and median



The result of moving the first of
Lord Rayleigh's seven values:
midmean



The result of moving the first of
Lord Rayleigh's seven values:
trimean



What if we move a different value in the same batch? Exhibit 12 shows some facts about the effects of moving each of Lord Rayleigh's seven values, one at a time. Panel B shows the greatest movement for each value. The mean greatest movements are about the same for the median and the two middle means, 8 to 9. Panel C shows how steep a change can occur, at least for small changes, while panel D shows the interval over which some change occurs. We can supply a partially comparable interval for the mean by asking how far any one value must move to move the mean seven units. The answer is $7 \cdot 7 = 49$.

In summary, then for this particular batch, we have the results in exhibit 13. On the whole, there is little doubt that the middle means outperform the mean and the median, except for special purposes.

We have looked at only one batch, of count 7. Would the results be similar for other batches of seven or more or less? Yes, so far as we have gone. (The reader is encouraged to invent and try out other examples.)

If we had to make up our minds between the midmean and trimean on the basis of this batch alone, we would, it would seem, prefer the midmean definitely but gently. We would not expect this clear but gentle preference to continue for batches of enough more observations to make the trimean arithmetically simpler.

exhibit 12 of chapter 4

The effects of moving each of Lord Rayleigh's seven values
(Unit = 0.00001, referred to 2.31000)

A) The GREATEST EFFECTS of MOVING ONE VALUE

	value moved	-14	1	10	17	24	28	30
Moved Lowest	Ordered values after move	low	low	low	low	low	low	low
		1	-14	-14	-14	-14	-14	-14
		10	10	1	1	1	1	1
		17	17	17	10	10	10	10
		24	24	24	24	17	17	17
		28	28	28	28	28	24	24
		30	30	30	30	30	30	28
	mean	low	low	low	low	low	low	low
	median	17	17	17	10	10	10	10
	midmean	17	17	14	11.7	9.3	9.3	9.3
	trimean	15.8	12	12	8.5	8.5	7.5	7.5
*		*	*	*	*	*	*	*
Moved Highest	Ordered values after move	1	-14	-14	-14	-14	-14	-14
		10	10	1	1	1	1	1
		17	17	17	10	10	10	10
		24	24	24	24	17	17	17
		28	28	28	28	28	24	24
		30	30	30	30	30	30	28
		high	high	high	high	high	high	high
	mean	high	high	high	high	high	high	high
	median	24	24	24	24	17	17	17
	midmean	23	23	23	20.7	18.3	17	17
	trimean	22	22	19.8	19.8	16.3	16.3	15.8

B) The LARGEST CHANGES

								(mean)
mean	great	great	great	great	great	great	great	(great
median	7	7	7	14	7	7	7	8
midmean	6	6	9	9	9	7.7	7.7	7.8
trimean	6.2	10	7.8	11.3	7.8	8.8	8.8	8.6

(Cont'd)

exhibit 12 (Cont'd)

C) The STEEPEST CHANGE of CENTRAL VALUE

mean	1/7	1/7	1/7	1/7	1/7	1/7	1/7	(1/7)
median	1	1	1	1	1	1	1	(1)
midmean	1/3	1/3	1/3	1/3	1/3	1/3	1/3	(1/3)
trimean	1/2	1/2	1/2	1/2	1/2	1/2	1/2	(1/2)

D) The INTERVAL over which CHANGE OCCURS

mean	all	all	all	all	all	all	all	(all)
median	{17 24	{17 24	{17 24	{10 24	{10 24	{10 17	{10 17	{(13) (23)
midmean	{10 28	{10 28	{1 28	{1 28	{1 28	{1 24	{1 24	{(4) (27)
trimean	{1 30	{-14 30	{-14 30	{-14 30	{-14 30	{-14 30	{-14 28	{(-12) (30)

exhibit 13 of chapter 4

Summary of behavior of Lord Rayleigh's seven values
When exactly one is moved

	mean	midmean	trimean	median
Mean greatest movement	Great	7.8	8.6	8
Greatest movement of all	Great	19	11.3	14
Steepest change	1/7	1/3	1/2	1
Interval of change	56*	23	42	8

* Only the beginning, but enough to make a movement of 8 in the mean, comparable with the greatest movements for the other central values.

4L. <u>Effect on pattern of residuals of choice of central value</u> (background)

We have thought about individual residuals. We have concluded, wisely, that we like residuals from middle means because each individual stray then has little effect except on its own residual. This is the big point. Its truth should dominate all other considerations.

There are, as we have already seen (for example, at panel D of exhibit 3), instances when it helps us to look, not at individual residuals, but at the pattern of all residuals from one or more batches. What effect does our choice of central value have on this pattern?

We have already seen that very stray observations will generate residuals of larger size from resistant central values. The mean, by contrast, will be more influenced by a very stray observation, thus reducing the size of the corresponding residual. (See panel B of exhibit 2 as an example.)

We ought to expect this effectto be balanced. The simplest balance would be for the residuals for observations quite near the central value to be smaller from resistant central values than from means. This is indeed the case. (The interested reader is asked to show how and why.)

What we expect from the pattern of residuals from a resistant value, then, is:

- more far-out values,
- more close-in values,
- fewer middle-distance values,

in comparison with the pattern of residuals from a mean.

Exhibit 14 starts an example. The "residuals from midmean" for the 12 batches of 7 monthly mean temperatures for New York City were shown in exhibit 3. To convert these into residuals from their means, it suffices to calculate

(residual) - (mean residual for batch)

as is done in panels A and B of exhibit 14. Exhibit 15 compares, with back-to-back stem-and-leaf, the two patterns of the 84 residuals. We see evidence of a considerable number of stray values in both stem-and-leaf displays. Clearly, however, both the number of apparent strays and their sizes are reduced when we take residuals from means. The total number of residuals from -1.0 to +1.0 is 20 for residuals from means and 30 for residuals from midmeans. Clearly both anticipated effects took place in this example -- there are both more strays and more small residuals.

The increase in the apparent number of strays seems usually to be a good thing. Having our attention drawn to the additional observations seems to help. The increase in the number of very small residuals does not seem to matter much.

If we were greatly concerned with patterns of residuals, we might even use residuals from resistant values where the residuals are large and residuals from means where the residuals are small. (The natural cross-over is near the hinges.)

exhibit 14 of chapter 4

The differing patterns of residuals from resistant central values and from means (from exhibit 3)

A) MEANS of RESIDUALS in PANEL C of EXHIBIT 3

Jan	Feb	Mar	Apr	May	Jun	July	Aug	Sept	Oct	Nov	Dec
2.2	0.4	2.1	0.5	0.4	-0.7	-0.1	-0.2	-0.5	0.3	0.0	-1.0

B) RESIDUALS FROM MEANS -- ordered within month

9.9	7.5	8.2	5.5	3.1	3.8	1.9	2.6	4.8	2.9	5.5	5.7
7.3	3.1	4.3	3.1	2.0	2.0	1.4	0.9	2.2	2.0	2.2	4.4
-1.9	1.5	-1.8	1.7	-0.1	0.9	1.1	0.8	1.6	0.5	1.3	2.7
-2.3	-1.1	-2.3	-1.1	-0.2	0.8	-0.2	0.5	0.3	-0.5	-0.4	1.9
-2.5	-1.5	-2.3	-2.1	-0.8	0.5	-0.5	-0.7	-0.4	-1.0	-0.9	-1.7
-3.1	-3.4	-2.4	-2.9	-1.7	-4.0	-0.9	-0.9	-2.0	-1.5	-2.7	-6.2
-7.1	-5.4	-3.6	-4.5	-2.2	-4.2	-2.7	-3.0	-6.6	-2.2	-4.8	-6.6

C) NOTE

Values found using:

Residuals from mean = residuals from any single value
less
mean of same residuals

exhibit 15 of chapter 4

Stem-and-leaf Comparison of Patterns of Residuals for the Same Data

from midmeans		from means	to	to
1	12		1	.
	11		.	.
3	10		2	.
5	9	9	3	1
	8	2	3	2
9	7	35	4	4
40	6		5	4
5	5	558	7	7
37	4	348	9	9
124556	3	1118	15	14
22344	2	00022679	20	21
0133377789	1	13456799	29	29
1223333 6789	0	34558899	to 40	to 38
9999776644333222221	-0	999875544221	✓ 19	✓ 11
986532110	-1	987755110	to 25	to 35
87754	-2	977543332210	16	25
20	-3	6410	11	13
99870	-4	8520	9	9
4	-5	8	4	6
	-6	662	3	4
631	-7	1	3	1

Source: 12 batches of 7 mean monthly temperatures.

Left side from panel C or D of exhibit 3.
Right side from panel B of exhibit 2.

4M. Background references

Some discussion of the mean will be found in almost any textbook of statistics. Most of these discussions will help you little with the point of view we are taking here. (Once you get this point of view clearly in mind, reading some of these discussions with a view to how you might want to change them could be profitable.)

Except that there will be fewer places where it is discussed, the same can be said for the median.

You will find almost nothing about middle means in print. (((Statisticians who have strayed into reading this introduction may be able to extract some nourishment from papers (i) by Crow and Siddiqui, at pages 353-389, and (ii) by Siddiqui and Raghunandanan at pages 950 to 953, both of volume 62 (1967) of the Journal of the American Statistical Association. They need to be warned that the aspect with which these writers are concerned is important, but is not our main issue.)))

*****

4P. Problems

This chapter invites two kinds of problems.

* problems for hand arithmetic *

Note: The annual rainfalls and winter snowfalls in New York City (problems 1-2,5-6) can be found in various sources, including The World Almanac.

(1) The snowfalls in New York City for 20 consecutive winters from 1918-19 to 1937-38 were (in inches) 3.5, 55.4, 18.2, 29.7, 55.2, 26.3, 27.9, 35.8, 21.9, 14.3, 13.3, 13.5, 9.7, 5.1, 24.5, 53.1, 29.0, 32.8, 11.9, 13.9. Find the trimean. Do you think other typical values would be very different?

(2) The snowfalls in New York City for 20 consecutive winters from 1938-39 to 1957-58 were (in inches) 31.9, 22.2, 35.0, 10.2, 27.6, 26.0, 26.7, 26.6, 33.2, 61.5, 43.0, 10.4, 10.9, 14.4, 9.1, 17.1, 10.9, 29.8, 19.1, 37.9. Find the trimean.

(3) Exhibit 16 gives the % of major party vote for Nixon and Goldwater in eight southwestern states in 1960 and 1964. Find the differences "Nixon-Goldwater" and then the mean, median, and midmean of these differences.

(4) During the winter of 1893-94, Lord Rayleigh was investigating the density of nitrogen from various sources. He had previously found indications of a discrepancy between the densities of nitrogen produced by removing the oxygen from air and nitrogen produced by decomposition of a chemical

exhibit 18 of chapter 4

% of major party vote for Republican Presidential candidates in eight southwestern states

	% for Nixon in 1960	% for Goldwater in 1964
Arizona	55.6%	50.5%
California	50.3%	40.8%
Colorado	54.9%	38.4%
Nevada	48.8%	36.1%
New Mexico	49.6%	40.6%
Oklahoma	59.0%	44.3%
Texas	50.5%	36.6%
Utah	54.8%	45.3%
(All U.S.)	(49.9%)	(38.7%)

From Richard M. Scammon 1965. America at the Polls: A Handbook of Presidential Election Statistics 1920-1964. Pittsburgh: University of Pittsburgh Press: 521 pp.

compound. The 1893-94 results established this difference with great definiteness and led him to investigate further the composition of air chemically freed of oxygen. This led to the discovery of argon, a new gaseous element.

Rayleigh's observations are summarized in exhibit 17. Calculate the midmean weight for "nitrogen" from air. Calculate the midmean weight for "nitrogen" prepared from either N_2O or NH_4NO_2. Compare and comment.

(5) In a study of cost of plastering in 1923 by the Bureau of Labor Statistics, plastering costs in dollars per square yard were obtained for 19 jobs in Atlanta, Georgia. These were 0.45, 0.16, 0.16, 0.15, 0.14, 0.15, 0.12, 0.29, 0.31, 0.20, 0.30, 0.27, 0.28, 0.40, 0.35, 0.14, 0.14, 0.18, 0.19. Calculate the trimean of these 19 values.

(6) Goulden gives the % of carotene in the flour and in the grain for 139 varieties of wheat. The percent carotene in the flour for the first 99 varieties was 2.39,3.11,2.15, 1.96,2.02,1.76,2.10,2.12,2.28,1.86,2.60,2.11,2.30,1.80,2.00, 2.05,2.09,2.33,2.29,2.30,1.97,2.36,1.73,1.72,1.70,1.63,1.93, 1.50,1.77,1.60,2.31,2.17,2.10,2.90,2.17,2.15,2.01,2.35,2.35, 2.00,2.18,2.47,2.25,1.77,1.68,2.46,1.86,1.71,1.93,1.81,1.89, 1.65,1.93,2.12,2.25,1.92,2.25,2.25,1.65,1.63,1.70,1.61,1.83, 1.60,1.37,1.96,1.88,1.92,1.89,1.99,1.82,2.12,2.16,2.14,1.63, 2.76,2.07,1.67,2.78,3.40,3.67,2.41,2.23,3.07,2.22,2.55,2.12, 1.94,1.95,1.59,2.00,1.77,1.98,1.97,1.97,1.83,2.00,1.96,2.00. Find the trimean.

exhibit 17 of chapter 4

Rayleigh's weights of a standard volume of "Nitrogen"

Date	Source of "nitrogen" Origin	Purifying agent	Weight
29 Nov. '93	NO	hot iron	2.30143
5 Dec. '93	"	"	2.29816
6 Dec. '93	"	"	2.30182
8 Dec. '93	"	"	2.29890
12 Dec. '93	Air		2.31017
14 Dec. '93	"	"	2.30986
19 Dec. '93	"	"	2.31010
22 Dec. '93	"	"	2.31001
26 Dec. '93	N_2O	"	2.29889
28 Dec. '93	"	"	2.29940
9 Jan. '94	NH_4NO_2	"	2.29849
13 Jan. '94	"	"	2.29889
27 Jan. '94	Air	Ferrous hydrate	2.31024
30 Jan. '94	"	"	2.31030
1 Feb. '94	"	"	2.31028

Source: Lord Rayleigh 1894. On an anomaly encountered in determinations of the density of nitrogen gas. Proc. Roy. Soc. (Lond.) 55, 340-344 (also in his Scientific Papers Vol. 4, pp. 104-108).

(From page 65 of C. H. Goulden 1939. Methods of Statistical Analysis. New York: John Wiley and Sons: 277 pp.)

(7) In 1950, Chu, Cummings, and Teixeira gave the results of drape tests in exhibit 18. Find residuals and central values.

(8) In 1948, Beckmann, McCullough, and Crane gave the results in exhibit 19 of microdeterminations of carbon monoxide in air. Choose a satisfactory central value. Find residuals and deviations. Combine across samples in a satisfactory way. Comment.

(9) In 1946, Johnson and Kline gave the results in exhibit 20 on fatty acids in dried egg powders. Find central values and residuals. What do you feel these results suggest?

(10) In 1950, Willard and Dean gave the results in exhibit 21 on the measurements of aluminum content. Find residuals and deviations. How should they be combined? Combine them and discuss the results.

(11) In 1950, Kirk, Kingsley, Kinsey, Sheehan and Swigert gave the results in exhibit 22 on the ultramicro measurement of nitrogen. Calculate and combine residuals.

(12) In 1950, Jackson and Beamish gave the results in exhibit 23 on the precipitation of platinum sulfide. Find and combine the residuals and deviations.

(13) In 1907±, Warren and Haviland gave the head measurements on termites in exhibit 24. Find and combine residuals.

(14) In 1948, Steinberg and Belic gave the lime-silica ratios in exhibit 25. Find and combine the residuals and deviations.

(15) In 1946, Smith gave the gossypol percentages in exhibit 26. Find and combine the residuals.

* theoretical problems *

(16) In a batch of eight values, what are the largest and smallest possible numbers of observations that can be greater than the mean? -- than the median? -- than the midmean? -- than the exact midmean? -- than the trimean? Can you prove your statements?

(17) Can you generalize the results of problem 21? How far?

(18) What light do you think (i) the conclusions and (ii) the detailed workings of problems 16 and 17 cast on behavior of means, medians, and middle means in the presence of wild values?

exhibit 18 of chapter 4

Consistency of drape measurements

A) DRAPE, measured by AREA in SQUARE INCHES

Sample A	Sample B
28.92	50.04
28.82	49.94
28.96	40.08
28.89	50.27
28.96	50.03
28.85	50.06
28.97	50.00
28.89	49.99
29.00	49.75
28.99	49.94

B) The SOURCE

C. C. Chu, C. L. Cummings, and N. A. Teixeira 1950. Mechanics of elastic performance of textile materials Part V. A study of factors affecting the drape of fabrics -- The development of a drape meter. Textile Research Journal 20: 539-548.

exhibit 19 of chapter 4

Microdetermination of carbon monoxide in air

A) PARTS PER MILLION of CARBON MONOXIDE

Sample	Standard*	Trial results**
A	90.5	95,96,92,102,103,93,101,92,95,90
B	184.6	184,202,215,204,195,201,201,169,182,192,
C	44.8	40,54,42,49,64,62,50,67,64,43,
D	320	261,279,281,278,269,264,266,261,266,276,
E	244.7	215,214,197,216,215,208,226,208,216,214,
F	25.8	26,23,25,25,21,22,27,27,21,25,
G	66.2	56,55,61,57,60,57,65,55,60,61,
H	137.8	128,119,119,123,117,122,127,121,122,119,
I	137.8	155,142,146,149,149,146,152,159,

* I_2O_5 method

** Beckmann-McCullough method, in order of analysis

B) The SOURCE

A. O. Beckmann, J. D. McCullough, R. A. Crane 1948. Microdetermination of carbon monoxide in air. Analytical Chemistry 20: 674-677.

exhibit 20 of chapter 4

Free fatty acids in dried egg powders

A) The FREE FATTY ACID VALUES of DRIED EGG POWDERS -- as oleic acid

Sample	Method	Determinations
A	Soxhlet extraction with ether	20|25788 21|02
A	Proposed	18|566 19|069 20|6
A	Proposed*	19|4566789 20|0
C	Soxhlet extraction with ether	3.8|2 3.9|4
C	Same*	3.4|7 3.5|247
C	Proposed	3.4|9 3.5|0225 3.6|35 3.7|3
C	Proposed*	3.3|7 3.4|77

* = adjusted to pH 4.0

B) The SOURCE

C. M. Johnson and L. Kline 1946. Determination of free fatty acids in dried egg powders. Industrial and Engineering Chemistry, Analytical Edition 18: 617-619. Table II on page 618.

exhibit 21 of chapter 4

Polarographic determination of aluminum

A) % Al_2O_3 -- oxygen-containing samples

Certified value	Sample	Determinations
4.16	Limestone	3|92,94, 4|16,36,
1.96	Silica brick	1.9|02336 2.0|04
1.91	Magnetite ore	1.71,1.75,1.77,1.90,2.10,
1.89	Soda-lime glass	1.59,1.61,2.14,2.14,
1.03	Iron ore	1.0|336799 1.1|236678
0.067	Dolomite	0.056,0.056

B) % Al_2O_3 -- alloys

1.13	Manganese bronze 62	1.1|444468
0.97	Manganese bronze 62b	0.8|9 0.9|4799
1.06	Nitralloy steel 106	.9|5 1.0|88 1.1|01
1.07	Nitralloy steel 106a	1.0|08 1.1|022
0.26	High silicon steel	.25|18 .26|009 .27|0

C) The SOURCE

H. H. Willard and J. A. Dean 1950. Polarographic determination of aluminum. Use of an organic reagent. Analytical Chemistry 22: 1264-1267. Table III on page 1266.

exhibit 22 of chapter 4

Kjeldahl ultramicrodetermination of nitrogen

A) The BATCHES -- microliters of 0.01N 4cl for a 10.42 microliter sample

Sample	Material	Determinations
A	Acetanilide	6|60,78,81, 7|03,57,60,72,
B	Acetanilide	6|13,37,39,46, 7|47,59,75,
C	$C_{14}H_{22}N_2O_2S$	6|25,63,67 7|32,43,63,
D	$C_{13}H_{20}N_2O_2S$	6|56,59,65,

B) The SOURCE

J. A. Kirk (??), A. Kingsley, D. Kinsey, F. Sheehan, and G. F. Swigert 1950. Kjeldahl ultramicrodeterminaion of nitrogen. Applications in the industrial laboratory. Analytical Chemistry 22: 604-611.

exhibit 23 of chapter 4

Precipitation of platinum sulfide by various methods

A) AMOUNTS of PRECIPITATE -- calculated as milligrams of platinum

Method	Platinum recovered from 10.12 mg Pt
Treadwell and Hall	10.2|89 10.3|13334
Gilchrist and Wichers	10.2|2334789
Hillebrand and Lundell	10.2|25 10.3|03688

B) The SOURCE

D. S. Jackson and F. E. Beamish 1950. Critical examination of platinum sulfide precipitation. Analytical Chemistry 22: 813-817.

exhibit **24** of chapter 4

Head breadths of termites

A) The DATA

Nest 668	**Nest 670**	**Nest 672**	**Nest 674**	**Nest 675**
	--- small soldiers ---			
227.3	24719	240.4	244.7	245.6
233.2	260.3	245.7	238.8	262.6
237.5	261.3	245.2	251.5	263.3
237.3	255.7	239.6	244.5	248.7
231.8	237.7	227.9	231.2	241.0
	--- large workers ---			
214.2	231.5	235.9	234.9	236.0
213.9	246.3	234.4	230.0	247.8
222.5	248.5	239.4	243.6	253.9
235.2	247.7	236.4	242.3	246.1
226.9	233.4	226.9	230.2	231.1

B) The SOURCE

Warren Haviland

Biometrika 6: 3?? - . page 333.

exhibit 25 of chapter 4

Lime-silica ratios by spark spectroscopy

A) The DATA

Sample of slag	Chemical value	Spectrographic values
A	0.67	0.70,0.68,0.63,0.68,
B	0.76	0.69,0.74,0.61,0.70,
C	1.85	2.0,1.7,1.9,1.7,
D	2.04	1.7,2.0,2.1,
E	2.27	2.3,2.3,2.2,
F	2.91	3.2,2.9,2.9,3.0,
G	3.70	3.4,3.9,3.6.3.6,
H	4.70	4.6,4.8,4.5,5.1,
I	6.00	6.6,5.9,5.8,6.2,
J	8.43	8.1,8.8,9.2,7.8,

B) The SOURCE

R. H. Steinberg and J. H. Belic 1968. Spark technique in spectrographic analysis of slags. Analytical Chemistry 20: 72-73. Table I on page 73.

exhibit 26 of chapter 4

Precision of determination of gossypol

A) The DATA

Sample of Cottonseed oil	Glavimetric % of gossypol
1	.0168
	.0177
	.0166
2	.0455
	.0542
	.0549
	.0518
	.0530
	.0524
3	.0555
	.0561
	.0542
	.0505
	.0508
	.0508

B) The SOURCE

F. H. Smith 1946. Spectrophotometric method for estimating gossypol in crude cottonseed oil. Industrial Engineering Chemistry, Analytical Edition 18: 41-43. Table I on page 42.